An analysis of GDP per capita

Jake Stimson, Michael Pitta, Michael Danylchuk

Department of Electrical Engineering, San Jose State University

Dr. Juzi Zhao

December 9, 2024

Abstract:

This study aims to analyze the impacts of an assortment of input variables related to GDP per capita for most of the world's countries (All countries included in the available public .csv data sets). The input variables covered will include Natural Resource Rents (% of GDP), the Human Capital Index, the Global Innovation Index, Worldwide Governance Indicators, and the Global Competitiveness Index. For more information on the statistical terminology used in this paper, please see the 'glossary' in the preceding sections for a more in-depth clarification.

Glossary:

Mean:

The mean measures a central tendency present in our data set. To calculate this mean, we can add all the numerical values in our data set and divide them by the total number of values.

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

In the above equation, x_i the values in the dataset are where i increments through each data position, and n is the full range of the data set.

Median:

The Median is also a measure of central tendency, representing the data set's positional center. Where outliers in the data set can heavily skew the mean, the median, taken from the mid-point of an ordered numerical set, can be more accurate regarding realistic central tendency.

If the data set contains an odd number of observations, the median can be found by picking the middle point of the data set. See the equation below:

$$\text{Median (odd)}: \frac{n+1}{2} \quad (1)$$

In the case that the data set contains an even number of observations, the median can be found by calculating the average of the two center numerical values. See the equation below:

$$\text{Median (even)}: \frac{\frac{n}{2} - \frac{n+1}{2}}{2} \quad (2)$$

In the above two equations, n represents the total number of data points in the data set.

Mode:

The mode refers to a statistical measure of frequency. More specifically, it measures the most frequently occurring value(s) that reside in the data set. There are typically two types of 'Mode' analysis: Single-mode or Multimodal. As their respective names suggest, single mode refers to the study of the most frequently occurring output in the data set. Multimodal analysis is used when multiple outputs occur more frequently than the rest.

To implement this analysis in software, we can iterate through our .csv data set and increment a counter variable that is associated with each individual output. Once the iteration is complete, we will essentially have a histogram that relates the number of occurrences in the data set to the individual outputs themselves.

Minimum:

The minimum represents the smallest output that exists in our data set. It essentially determines the low boundary for our data set.

Maximum:

The maximum represents the largest output that exists in our data set. It essentially determines the high boundary of our data set.

Variance:

The variance gives us a good measure of the overall spread of our data as it relates to the mean. It is directly related to the standard deviation, and it generally gives us a good idea of how far a number deviates from the average (mean).

$$Variance = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \quad (3)$$

Here, x_i represents each data point in the set and $\bar{x}$ represents the mean of the set. N is the number of samples in the set.

Standard Deviation:

The standard deviation will provide a quantitative value for the dispersion of the set of values. This gives us an easy way to get an average distance that a value lies from the mean.

$$\sigma = \sqrt{Variance} \tag{4}$$

Scatter Plot:

The scatter-plot gives us a graphical view of a series of points in 2-dimensional space. One variable will be plotted on the x-axis, and the other is plotted on the y-axis. This representation should give us a visual idea of how changing an input variable affects the output. If there is an obvious correlation, the scatter data will follow a general curve.

Matrix:

A matrix allows us to store a variation of numerical values inside a 2-dimensional grid. This way, we can quickly use software to iterate through the matrix (array) and compute transformations based on the data. For our specific case, the matrix that represents our data sets will include GDP data for each country going down the first column, followed by each of our five input variables and their respective outputs tied to their own column.

Correlation Coefficient:

This is a numerical measure of the strength of correlation with respect to a linear model. Denoted by r, the closer this value is to 1, the relationship describes a perfect positive linear relationship - inversely, the closer r is to -1 indicates a perfect inverse linear relationship. To solve for the Correlation Coefficient, see equation 5 below.

$$r = \frac{Cov[X, Y]}{\sqrt{Var(X) \times Var(Y)}} \tag{5}$$

Feature Scaling:

The idea of feature scaling is to standardize the numerical values that exist within our data set. This allows us to easily view relationships between the data as they are all normalized to the same scale. In our case, the data will be scaled between 0 and 1.

Linear Regression Models:

In order to extrapolate data that we can see has a linear relationship, we can use a linear regression model. This model allows it to predict a dependent variable based on the input of an independent variable. For a simple linear regression between one independent and one dependent variable, equation 6 below can be used.

$$Y = \beta_0 + \beta_1 X \tag{6}$$

Where 'Y' is the dependent output variable, and 'X' is the independent input variable.

Mean Squared Error (MSE):

Mean square error is a measure of the average squared difference between the actual values and the predicted values in a regression model. A lower MSE indicates a closer fit of the model to the data, which suggests high accuracy. A higher MSE indicates larger errors and less accurate predictions

$$MSE = \frac{\sum_{j=1}^{m} (actual\ value - predicted\ value)^2}{m} \quad (8)$$

R-Squared (R^2):

The coefficient of determination quantifies the proportion of variance in the dependent variable explained by the independent variables. Ranges from 0 (no explanatory power) to 1 (perfect fit).

Natural Resource Rents:

Percentage of GDP derived from natural resources. High values may indicate resource dependence.

Human Capital Index (HCI):

A composite index measuring a country's education and health outcomes is used to evaluate workforce productivity.

Global Innovation Index (GII):

An index assesses a country's innovation capabilities, including R&D investments and technological output.

World Governance Indicators (WGI):

Measures governance quality based on political stability, regulatory quality, and government effectiveness.

Global Competitiveness Index (GCI):

Evaluates infrastructure, macroeconomic stability, and innovation to measure a nation's competitiveness.

GDP per Capita:

Gross domestic product is divided by the population, serving as an indicator of economic productivity and living standards.

Figures:

Table 1: GDP (Gross domestic product) total	
Mean	2420333704851.444
Median	48065653179.0
Mode	1522011045163.0
Minimum	41629064.22
Maximum	76588030455296.0

Variance	6.915286640586605e+25
Standard Deviation	8315820248530.2705

Table 2: Natural Resources Rents (% of GDP)	
Mean	4.874227604214944
Median	2.249230874
Mode	1.51520453728299
Minimum	0.0002523353870072
Maximum	44.5556159738805
Variance	42.27538413370923
Standard Deviation	6.501952332469782

Table 3: Human Capital Index	
Mean	0.5694691270382696
Median	0.5745359

Mode	0.369
Minimum	0.28607464
Maximum	0.88708365
Variance	0.021189933213148084
Standard Deviation	0.14556762419284064

Table 4: Global Competitiveness Index (World Economic Forum 2016)	
Mean	705.4798181803242
Median	41.0
Mode	1137.8593991660255
Minimum	-39.7
Maximum	14264600.0
Variance	2350922183.0729036
Standard Deviation	14190.808162056075

Table 5: Global Innovation Index (2016)	
Mean	50.098474665612436
Median	43.4
Mode	1.0
Minimum	0.1
Maximum	129.0
Variance	1481.1941211536594
Standard Deviation	38.486284844781515

Table 6: World Governance (Government Stability)	
Mean	0.014429500866272135
Median	0.1200919598
Mode	1.169785261
Minimum	-3.312951088

Maximum	1.964210629
Variance	0.9710498096734878
Standard Deviation	0.9854185961678863

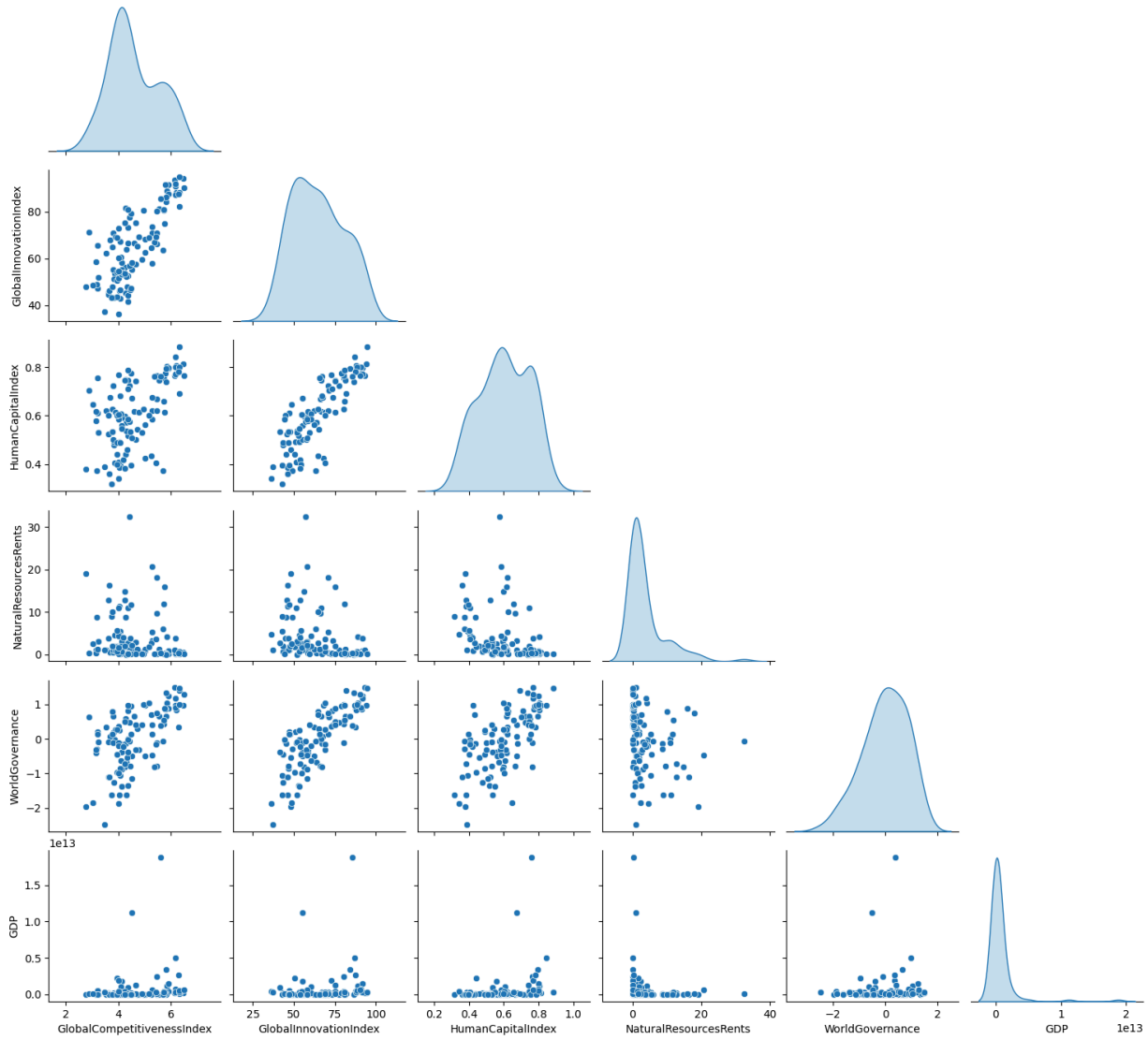

Figure 1: Scatter plot matrix generated using data prior to normalization.

```
Correlation Coefficients
========================

                          GlobalCompetitivenessIndex  GlobalInnovationIndex  HumanCapitalIndex  NaturalResourcesRents  WorldGovernance       GDP
GlobalCompetitivenessIndex                    1.000000               0.764186           0.540679              -0.128816         0.595815  0.192206
GlobalInnovationIndex                         0.764186               1.000000           0.836547              -0.304071         0.799905  0.181880
HumanCapitalIndex                             0.540679               0.836547           1.000000              -0.373828         0.653651  0.234265
NaturalResourcesRents                        -0.128816              -0.304071          -0.373828               1.000000        -0.255129 -0.133715
WorldGovernance                               0.595815               0.799905           0.653651              -0.255129         1.000000  0.059764
GDP                                           0.192206               0.181880           0.234265              -0.133715         0.059764  1.000000

Key Observations:
- Correlation coefficients range from -1 (strong negative) to 1 (strong positive).
- Identify the strongest relationships for modeling.
```

Figure 2: Correlation coefficients generated after data normalization.

Introduction and Background:

This project is a display of how statistical techniques can be implemented in service of economic analysis. Many organizations, industries, and educational fields have been studying and predicting GDP per capita across various regions for years. By compiling and creating models based on historical data, researchers can improve predictive modeling, optimize policy recommendations, and assess economic risk. In contrast to the wide array of data and applications across different domains, this report seeks to narrow the focus to GDP per capita with five potential indicators to trends.

GDP per capita is an important metric for understanding both the economic health and living standards of the population of the world's nations. Consequently, economists and lawmakers commission sophisticated analyses to monitor trends, predict future results, and craft strategies to direct the predictable growth of the economy. This goal is achieved by examining the relationship between both dependent and independent input variables like natural resources and education, as well as the resulting GDP per capita. The results from the analysis assist in both making accurate economic predictions and interventions when needed.

This project focused on independent variables and explores the relationship between each and the GDP per capita. This was done by the application of statistical methods and regression analysis with the goal of achieving a greater understanding of economic growth drivers.

Discussion of Data:

The goal outlined in this project is to predict GDP per capita based on five independent factors. The output variable is GDP per capita, which is a widely known indicator of economic growth and living standards. Through the analysis of the relationship between GDP per capita and input factors, a more clear understanding of economic drivers can be formed. The input variables in this report will include natural resources, human capital (education and skill level), technological advancement, government policies and economic stability, and infrastructure. Each of these factors has a significant impact on a country's GDP per capita due to their direct influence on productivity, economic growth, and prosperity. The following section outlines the input variables and a selection of their most significant effects on GDP per capita.

Natural Resources: Countries rich in natural resources, such as oil, gas, minerals, and forests, often experience higher GDP per capita, as the extraction and export of these resources contribute significantly to national income. The role of natural resource rents as a percentage of GDP is important in evaluating the economic strength derived from these resources.

Human Capital (Education and Skill Level): A well-educated and healthy workforce is a key determinant of GDP per capita. Countries that invest in education and healthcare systems tend to

have higher productivity, driving economic growth. The Human Capital Index serves as a measure of this potential by considering factors such as education levels and health outcomes.

Technological Advancement: Countries that lead in technological innovation often experience higher GDP per capita. The Global Innovation Index assesses a country's innovation capacity, which directly affects economic output by allowing new industries, improving productivity, and enhancing competitiveness in global markets.

Government Policies and Economic Stability: Strong governance and effective economic policies provide the foundation for a stable and prosperous economy. Government effectiveness, political stability, and regulatory quality can all contribute to higher GDP per capita by ensuring efficient resource allocation, fostering business growth, and maintaining investor confidence.

Infrastructure: The quality of a country's infrastructure, as measured by the Global Competitiveness Index, is another critical factor in economic performance. Efficient transport systems, reliable energy grids, and advanced communication networks enable higher productivity and economic growth, thereby improving GDP per capita.

Now that the input variables have been properly defined and their effects on GDP per capita have been introduced, it’s time to examine the dataset associated with these factors. In the dataset, each row represents one country's data, with information on the variables listed above. The dataset contains more than 100 entries for each variable across numerous countries, providing a

large foundation for analysis. By applying statistical techniques in spreadsheets, we can calculate the summary statistics for each of the input variables.

Table 7: Summary of Results Normalized

Input	Mean	Median	Mode	Min	Max	Variance	Standard Deviation
Natural Resource Rents	0.1003	0.0461	N/A	0.0	1.0	0.0217	0.1474
Human Capital Index	0.4639	0.4552	0.1286	0.0	1.0	0.0654	0.2557
Global Competitiveness Index	0.5780	0.5374	0.5453	0.0	1.0	0.0378	0.1945
Global Innovation Index	0.5963	0.5640	0.9075	0.0	1.0	0.0436	0.2088

World Governance	0.6305	0.6505	0.8495	0.0	1.0	0.8118	0.8145
GDP per Capita	0.0372	0.0002164	N/A	0.0	1.0	0.0234	0.1530

After the results were compiled into Table 7 using the Python code found in the appendix, the statistical results were reviewed to determine any outliers within the dataset. Datapoints that are deemed outliers significantly differ from other observations and will skew the results. The result is that the removal of outliers is important to ensure that a proper analysis can be made to determine any relationship between the input variables and GDP per capita. The resulting cleaned-up data can then be analyzed for any correlation between the input variables and GDP.

The next step was to create multiple scatter plots, resulting in a matrix of plots. This technique allows a visual representation of the data distribution, and the resulting overlapping graphs allow an easy-to-digest representation of the five input variables' trends towards correlation, or lack thereof. The greatest correlation between GDP and our selected variables was the Human Capital Index.

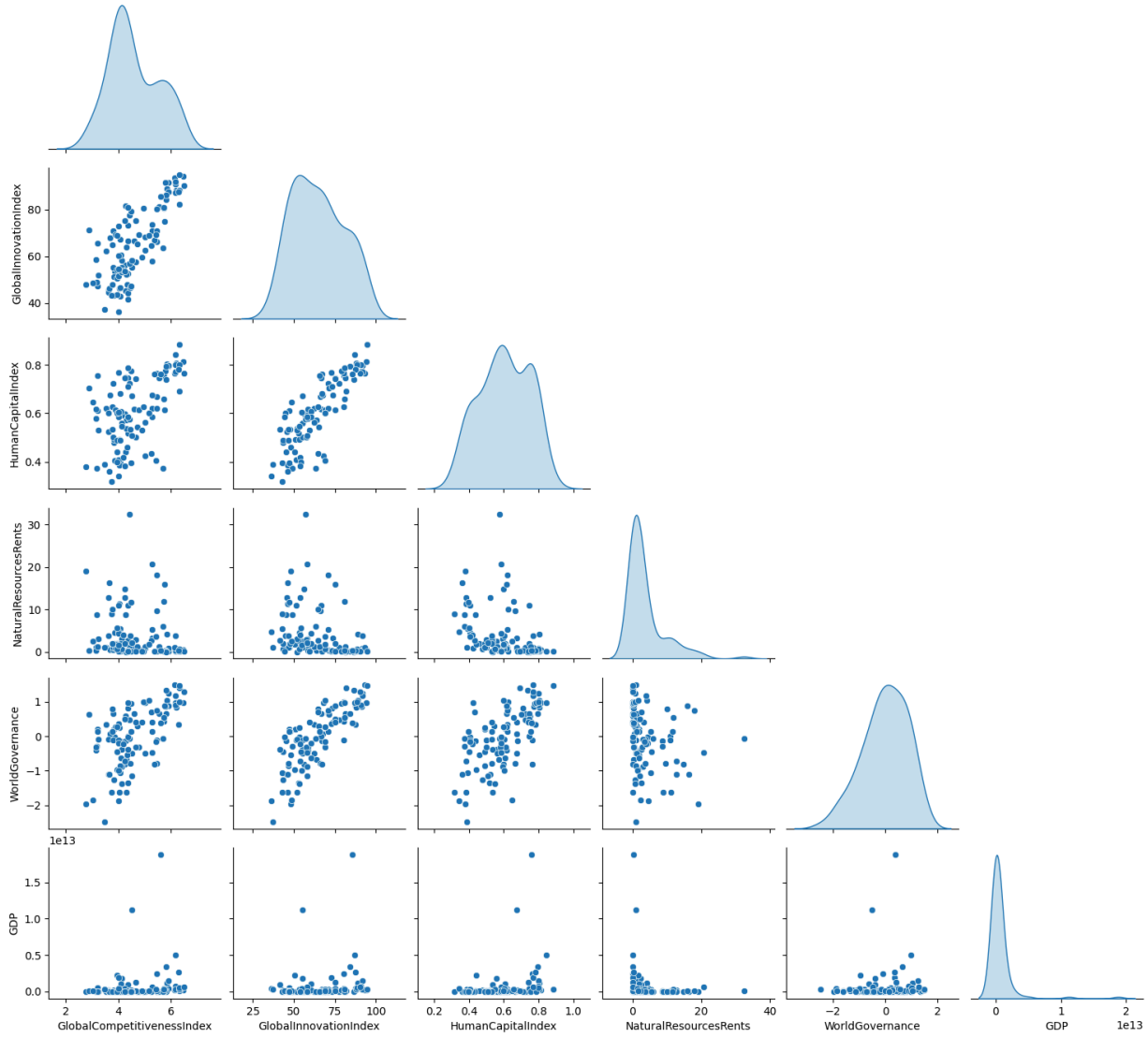

Figure 3: Scatter plot matrix generated by the five input variables data.

A quick review of the scatter plot matrix reveals that there is a stronger correlation between some variables and GDP when compared to others. Although none of the variables have a non-zero correlation, we can make some observations as to what the strongest correlations are.

Table 8: Correlation Coefficients						
	Global Competitiveness Index	Global Innovation Index	Human Capital Index	Natural Resource Rents	World Governance	GDP
Global Competitiveness Index	1.0	0.764186	0.540679	-0.128816	0.595815	0.192206
Global Innovation Index	0.764186	1.0	0.836547	-0.304071	0.799905	0.181880
Human Capital Index	0.540679	0.836547	1.0	-0.373828	0.653651	0.234265
Natural Resource Rents	-0.128816	-0.304071	-0.373828	1.0	-0.255129	-0.133715
World Governance	0.595815	0.799905	0.653651	-0.255129	1.0	0.059764
GDP	0.192206	0.181880	0.234265	-0.133715	0.059764	1.0

Analysis:

$$Output\ Variable\ = \Sigma^{n}_{k=1}\ constant_{k} * input\ variable\ _{k} + constant_{n+1} \quad (9)$$

The equation for output variables was used in the tables below to determine the linear n=1-3, and equation 8 in the glossary was then used to calculate the MSE.

Table 9: Linear Model Output N=1

Models	Coefficients	Intercept	MSE	R^2
GCI	0.01687566	0.019200235912278547	0.00590063463180809	-0.1323894254686555
GII	0.01582294	0.019673167015119113	0.00581952389709207	-0.1168234757526105
HCI	0.01652	0.021535306800049128	0.00586119450134794	-0.1248204455403789

Reviewing Table 9 with models using GCI, GII, and HACI to predict GDP shows lackluster performance with negative R^2 values, indicating explain less than a means-based prediction. The MSE of each model being around 0.0058 indicates that the predictions are comparable, but not with high accuracy. The coefficients for all variables are positive which suggests that there's a relationship with GDP, but small magnitudes indicate weak influence. The biggest limitation is

the negative R^2 values, but this may be due to missing data or oversimplification. Improving data quality, incorporating additional variables, and exploring nonlinear models are recommended steps to enhance predictive performance.

Table 10: Linear Model Output N=2

Models	Coefficients	Intercept	MSE	R^2
GCI, GII	0.01164111, 0.00899798	0.016921753572956465	0.0058666640956151735	
GCI, HCI	0.00829155, 0.01327677	0.018031775608982614	0.005883689229401015	
GII, HCI	0.00471461, 0.0141875	0.019654391773786146	0.005850323473633742	

The models found in Table 10 have a similar issue with negative R^2 values which means that they have a difficult time making accurate predictions. The GII and HCI model performs slightly better, with the lowest MSE (0.00585) and least negative R^2 (-0.12273). All coefficients are positive, reflecting a direct relationship between the variables and GDP, but their small magnitudes suggest weak impacts. These results highlight the need for further improvements, such as including all three variables in a combined model, enhancing data quality, or exploring non-linear modeling approaches to better capture relationships.

Table 11: Linear Model Output N=3				
Models	Coefficients	Intercept	MSE	R^2
GCI, GII, HCI	0.0078746, 0.00103562, 0.01298983	0.017794788677 65881	0.005880103392 232157	

The model combining all three variables exhibits limited predictive power, with a negative R^2 value (-0.1284) indicating it could be better with predictions. The coefficients reveal a weak relationship between GDP and the input variables, with GII (0.00104) contributing the least and HCI (0.01299) the most. The MSE (0.00588) is slightly better than that of the two-variable models but shows minimal improvement. These results highlight the insufficiency of the input variables in explaining GDP, suggesting the need for data exploration, inclusion of additional predictors, and consideration of nonlinear models for improved performance.

After the variables were generated a heuristic model was generated using the following equation, and the results can be seen in Table 12.

$$Output\ Variable\ =\ f(input\ variable1,\ input\ variable2.\ input\ variable3 \quad (10)$$

Table 12: Heuristic Model	
f(GCI, GII, HCI) = a*GCI^2 + b*GII + c*HCI^2 +d*GCI*GII + e	
MSE: 0.005210046438719207	
a	-0.000953710577931045

b	-0.00043324368493157656
c	0.008886028136608203
d	-0.0003043811870424573
e	0.05308882885591878

The heuristic model for predicting GDP, given by the equation found in Table 12 includes quadratic terms (GCI^2, HCI^2), an interaction term (GCI * GII), and a constant. The optimized coefficients for the model are a-e in Table 12, and the model's MSE can be found in the table. The most influential term in the model is HCI^2, which has the largest positive coefficient, suggesting a significant positive effect on GDP. In contrast, the negative coefficients for GCI^2, GII, and the interaction term indicate weaker or slightly adverse effects on GDP. Compared to previous linear models, the heuristic model demonstrates better predictive performance, with a lower MSE. Despite its increased complexity due to the quadratic and interaction terms, the model's performance justifies this complexity. The model's results underscore the importance of human capital in driving economic growth, suggesting that investments in human capital are crucial. Recommendations include continuing to use the heuristic model for GDP predictions, exploring additional terms to improve MSE further, validating the model on new data, and refining the understanding of HCI's role in GDP growth.

Results:

Scatter Plot Matrix and Correlation Analysis:

The scatter plot matrix provided a visual representation of relationships between GDP and the five independent variables. Key patterns and trends included:

1) Human Capital Index (HCI): HCI had the strongest positive correlation with GDP (r = 0.234), though the relationship remained weak overall. The scatter plot suggests that higher HCI values might be associated with higher GDP, but the trend is not strongly linear.
2) Global Competitiveness Index (GCI) and Global Innovation Index (GII): These variables also displayed weak positive correlations with GDP (r = 0.192 and r = 0.182, respectively). Their scatter plots showed clustering but lacked a clear, consistent trend, highlighting the limitations of linear relationships in explaining GDP variations.
3) Natural Resource Rents (NRR): The small negative correlation (r = -0.134) suggests that countries more reliant on natural resources may experience slightly suppressed GDP levels, potentially due to resource dependency challenges.
4) World Governance Indicators (WGI): With almost no correlation (r = 0.060), WGI did not appear to have a significant relationship with GDP. Its scatter plot lacked discernible patterns, making it less relevant for modeling.

These findings point to the complexity of GDP as a dependent variable and suggest that relationships with these independent variables may involve non-linear or interaction effects.

Linear Models

The performance of the linear models highlighted the difficulty of using the given variables to predict GDP effectively:

Single-Variable Models (N1):

1) Models using one input variable (e.g., GCI, GII, or HCI) showed weak predictive power, with R^2 values consistently negative. This indicates that these models performed worse than simply predicting the mean GDP.
2) The Mean Squared Errors (MSEs) ranged from 0.0058 to 0.0059, showing minimal variation across models.

Two-Variable Models (N2):

1) Pairing variables (e.g., GCI + GII, GCI + HCI) provided slight improvements in MSE compared to single-variable models. However, negative R^2 values persisted, reflecting poor model fit.
2) The combination of GII and HCI achieved the best performance among N2 models, with the lowest MSE (0.00585), but the improvement was marginal.

Three-Variable Model (N3):

1) Incorporating three variables (GCI, GII, HCI) yielded an MSE of 0.00588, comparable to the best N2 model. The negative R^2 value (-0.128) suggests the added complexity did not enhance the model's ability to explain GDP variability.

The results indicate that linear models alone are insufficient for capturing the underlying relationships, as the chosen variables explain very little of the variation in GDP.

Heuristic Model

The heuristic model demonstrated significant improvement over linear models. By incorporating non-linear terms and interactions, it achieved an MSE of 0.00521, the best among all models. Key findings include:

1) The quadratic term for HCI (HCI^2) emerged as the most impactful, with the highest positive contribution to GDP. This highlights the role of human capital as a critical driver of economic growth.
2) Interaction effects, such as GCI * GII, and higher-order terms captured relationships missed by linear models. These findings suggest that GDP is influenced by complex, non-linear interactions between variables.
3) Despite the improvement, the heuristic model's predictive power remains limited, indicating that additional variables or transformations might be needed for further enhancements.

Challenges and Limitations

1) Weak Correlations: The weak correlations between GDP and independent variables limited the effectiveness of both linear and heuristic models. This underscores the need for better predictors or additional variables.
2) Multicollinearity: The strong correlations between some input variables (e.g., GII and HCI, $r = 0.837$) suggest multicollinearity, which may have affected model performance. Regularization techniques like Ridge or Lasso regression could mitigate this issue.
3) Variable Selection: WGI and NRR contributed little to explaining GDP variability. Excluding these variables in future models could simplify the analysis without sacrificing accuracy.

Implications of Results

The findings provide valuable insights into the factors influencing GDP:

1) Human Capital Matters: The consistent importance of HCI across models highlights the role of education and workforce productivity in economic development. Policymakers should prioritize investments in human capital to drive GDP growth.
2) Complex Relationships: The improved performance of the heuristic model indicates that GDP is influenced by non-linear and interaction effects. Future models should account for these complexities to achieve better predictive power.
3) Diversification is Key: The negative impact of NRR on GDP aligns with the "resource curse" hypothesis, emphasizing the importance of diversifying economies beyond natural resource reliance.

Future Directions

The results suggest several avenues for further analysis:

1) Expand Data: Including additional economic indicators (e.g., trade balance, industrial output, or population) could improve model performance.
2) Explore Advanced Techniques: Non-linear machine learning methods like random forests or neural networks could capture hidden patterns and relationships in the data.
3) Validate Models: Testing the heuristic model on new datasets is essential to assess its generalizability and robustness.

By addressing these challenges and exploring new directions, future analyses can build on the insights gained from this project to better understand and predict GDP growth.

Conclusion:

This project provided a comprehensive exploration of the relationships between GDP per capita and five key variables: Global Competitiveness Index (GCI), Global Innovation Index (GII), Human Capital Index (HCI), Natural Resource Rents (NRR), and World Governance Indicators (WGI). By leveraging statistical techniques, linear regression, and heuristic modeling, we uncovered insights into economic growth drivers and revealed the complexities of modeling GDP.

The analysis highlighted that while certain variables, such as HCI, demonstrated a stronger connection to GDP, overall correlations remained weak, suggesting the need for more nuanced models. The scatter plot matrix and correlation coefficients revealed the multifaceted nature of the relationships, underscoring the importance of non-linear interactions and higher-order effects. Linear regression models offered limited predictive power, as evidenced by consistently negative values. These results indicated that the chosen variables alone could not adequately explain GDP variability. However, the heuristic model, incorporating quadratic terms and interaction effects, achieved a lower Mean Squared Error (MSE) of 0.00521, demonstrating the value of capturing non-linear relationships.

Key challenges included weak correlations, potential multicollinearity among independent variables, and the limited scope of input factors. These challenges highlighted the complexity of economic phenomena and the need for advanced modeling techniques. Despite these limitations, the consistent importance of HCI across all models reinforced the critical role of human capital in driving economic growth.

Looking forward, this study lays the groundwork for future analyses by emphasizing the need for:xw

- Expanding datasets to include additional economic indicators.
- Employing advanced machine learning techniques to uncover hidden patterns.
- Validating models with new data to ensure robustness and generalizability.

Ultimately, this project serves as a valuable demonstration of how statistical and computational methods can provide insights into economic dynamics, guiding both researchers and policymakers toward a better understanding of the factors influencing GDP per capita. Through continued exploration and refinement, these methods hold the potential to support more accurate predictions and informed decision-making.

Discussion of Contribution:

Jake Stimson: Research, data analysis/scripting, worked on glossary

Michael Pitta: Wore analysis, data tables, figures, references

Michael Danylchuk: Built Git Repository/Scripts for Steps 3-8, generated scatter plot matrix, generated linear and heuristic models. Helped with data tables and worked on conclusion

References

Danylchuk, M. (n.d.). Mikedan37/ee102_engineeringstatistics_finalproject: Final project for EE 102 at Sjsu. GitHub. *https://github.com/Mikedan37/EE102_EngineeringStatistics_FinalProject*

Global Competitiveness Report 2020. World Economic Forum. (2020, December 16). *https://www.weforum.org/reports/the-global-competitiveness-report-2020*

Global innovation index. (2024). *https://www.globalinnovationindex.org/Home*

World Bank Group. (2024). Databank. World Bank. *https://databank.worldbank.org/*

Yates, R. D., & Goodman, D. J. (2015). Probability and stochastic processes: A friendly introduction for electrical and computer engineers. Wiley.

Zhao, J. (2024a). EE102 Project. EE 102, 1–6.

Code Appendix:

https://github.com/Mikedan37/EE102_EngineeringStatistics_FinalProject

```python
def analyze_natural_resources_data(file_path):
    # Read the CSV file, assuming data starts from row 2 and column 1 is country names and column 2 is data
    df = pd.read_csv(file_path, skiprows=1)

    # Check the first few rows of the dataframe to understand its structure
    print("First few rows of the data:\n", df.head())

    # Extract the 'Country Name' and 'Natural Resources per Capita' columns
    country_names = df.iloc[:, 0]  # Column 1: Country names
    resources_per_capita = df.iloc[:, 1]  # Column 2: Natural resources per capita

    # Convert the 'Natural Resources per Capita' to numeric, coercing errors to NaN
    resources_per_capita = pd.to_numeric(resources_per_capita, errors='coerce')

    # Calculate the statistical measures
    mean_value = resources_per_capita.mean()
    median_value = resources_per_capita.median()
    mode_value = resources_per_capita.mode()[0]  # Getting the first mode
    std_dev = resources_per_capita.std()
    variance = resources_per_capita.var()

    # Covariance is typically calculated between two variables, but since we only have one variable here,
    # we'll compute the covariance of the data with itself, which should be equal to the variance.
    covariance = resources_per_capita.cov(resources_per_capita)

    # Print the results
    print("\nStatistical Analysis of Natural Resources per Capita:")
    print(f"Mean: {mean_value}")
    print(f"Median: {median_value}")
    print(f"Mode: {mode_value}")
    print(f"Standard Deviation: {std_dev}")
    print(f"Variance: {variance}")
    print(f"Covariance: {covariance}")
```

Figure 4: Function written to import the Natural Resources CSV file data and perform statistical analysis.

```python
def analyze_gdp_data(file_path):
    # Read the CSV file, skipping the first row to ensure row 2 starts from index 0
    df = pd.read_csv(file_path, skiprows=1)


    print("First few rows of the data:\n", df.head())

    country_names = df.iloc[:, 3]  # Column 5: Country names (index 3)
    gdp_values = df.iloc[:, 4]     # Column 6: GDP (index 4)

    gdp_values = pd.to_numeric(gdp_values, errors='coerce')

    mean_value = gdp_values.mean()
    median_value = gdp_values.median()
    mode_value = gdp_values.mode()[0]  # Getting the first mode
    std_dev = gdp_values.std()
    variance = gdp_values.var()

    covariance = gdp_values.cov(gdp_values)

    # Print the results
    print("\nStatistical Analysis of GDP:")
    print(f"Mean: {mean_value}")
    print(f"Median: {median_value}")
    print(f"Mode: {mode_value}")
    print(f"Standard Deviation: {std_dev}")
    print(f"Variance: {variance}")
    print(f"Covariance: {covariance}")
```

Figure 5: Function written to import the GDP CSV file data and perform statistical analysis.

```python
def normalize_and_plot(file_path):
    # Read the CSV file, skipping the first row to ensure row 2 starts from index 0
    df = pd.read_csv(file_path, skiprows=1)

    # Print the first few rows to inspect the data structure
    print("DataFrame head:\n", df.head())

    # Check the column names to ensure correct indexing
    print("\nColumn names:\n", df.columns)

    # Extract the relevant columns: Country Name, Natural Resources per Capita, and GDP
    country_names = df.iloc[:, 3]  # Column 5: Country names (index 4)
    resources_per_capita = pd.to_numeric(df.iloc[:, 1], errors='coerce')  # Column 2: Natural resources per capita (index 1)
    gdp_values = pd.to_numeric(df.iloc[:, 4], errors='coerce')  # Column 6: GDP (index 5)

    # Normalize the Natural Resources per Capita data between 0 and 1, excluding zeros
    resources_per_capita_non_zero = resources_per_capita[resources_per_capita > 0]
    min_resources_non_zero = resources_per_capita_non_zero.min()  # Find the smallest non-zero value
    normalized_resources = (resources_per_capita - min_resources_non_zero) / (resources_per_capita.max() - min_resources_non_zero)

    # Normalize the GDP data between 0 and 1, excluding zeros
    gdp_values_non_zero = gdp_values[gdp_values > 0]
    min_gdp_non_zero = gdp_values_non_zero.min()  # Find the smallest non-zero value
    normalized_gdp = (gdp_values - min_gdp_non_zero) / (gdp_values.max() - min_gdp_non_zero)

    # Plotting the normalized data
    plt.figure(figsize=(10, 6))
    plt.scatter(normalized_resources, normalized_gdp, color='b', label="Countries")

    # Adding labels and title
    plt.title('Normalized Natural Resources vs Normalized GDP', fontsize=14)
    plt.xlabel('Normalized Natural Resources per Capita', fontsize=12)
    plt.ylabel('Normalized GDP', fontsize=12)

    # No annotation (country names) is added in this version of the code

    # Show the plot
    plt.tight_layout()
    plt.show()
```

Figure 6: Function written to normalize GDP and natural resources data and generate a scatter plot with Natural Resources per capita on the x-axis and GDP per capita on the y-axis.

```python
import pandas as pd

# Load the data
file_path = '/Users/mdanylchuk/HCI_data.csv'
data = pd.read_csv(file_path)

# Clean the data by focusing on the HCI column and removing missing values
hci_cleaned = data["Human capital index (HCI) (scale 0-1)"].dropna()

# Compute the required statistics
mean_hci = hci_cleaned.mean()
median_hci = hci_cleaned.median()
mode_hci = hci_cleaned.mode().iloc[0] if not hci_cleaned.mode().empty else None
min_hci = hci_cleaned.min()
max_hci = hci_cleaned.max()
variance_hci = hci_cleaned.var()
std_dev_hci = hci_cleaned.std()

# Print the results
statistics = {
    "Mean": mean_hci,
    "Median": median_hci,
    "Mode": mode_hci,
    "Minimum": min_hci,
    "Maximum": max_hci,
    "Variance": variance_hci,
    "Standard Deviation": std_dev_hci,
}

for stat, value in statistics.items():
    print(f"{stat}: {value}")
import pandas as pd

# Define the file path
file_path = '/Users/mdanylchuk/HCI_data.csv'

# Load the dataset
data = pd.read_csv(file_path)

# Filter necessary columns and drop rows with missing 'HCI' values
filtered_data = data[['Entity', 'Human capital index (HCI) (scale 0-1)']].dropna()
filtered_data.rename(columns={'Human capital index (HCI) (scale 0-1)': 'HCI'}, inplace=True)

# Group by 'Entity' (Country) and calculate statistics
summary_stats = filtered_data.groupby('Entity')['HCI'].agg(
    Mean='mean',
    Median='median',
    Mode=lambda x: x.mode().iloc[0] if not x.mode().empty else None,
    Max='max',
    Min='min'
).reset_index()

# Save the cleaned and summarized data to a new CSV file
output_path = '/Users/mdanylchuk/HCI_summary_stats.csv'
summary_stats.to_csv(output_path, index=False)

print(f"Summary statistics saved to {output_path}")
```

Figure 7: Script to sort the Human Capital Index CSV file.

```python
import pandas as pd

# Load the dataset
file_path = '/Users/mdanylchuk/Downloads/GlobalCompetitivenessIndex.csv'
data = pd.read_csv(file_path)

# Step 1: Extract rows and columns starting from the relevant data (row 3 and column "11H")
# Assuming relevant data starts at the 11th column (adjust based on actual structure)
numeric_data = data.iloc[3:, 10:]

# Step 2: Convert all values to numeric, coercing errors to NaN
numeric_data = numeric_data.apply(pd.to_numeric, errors='coerce')

# Step 3: Drop rows and columns that are entirely NaN
cleaned_data = numeric_data.dropna(how='all', axis=0).dropna(how='all', axis=1)

# Step 4: Compute statistics for the cleaned dataset
statistics = {
    "Mean": cleaned_data.mean().mean(),
    "Median": cleaned_data.median().median(),
    "Mode": cleaned_data.mode().iloc[0].mean() if not cleaned_data.mode().empty else None,
    "Minimum": cleaned_data.min().min(),
    "Maximum": cleaned_data.max().max(),
    "Variance": cleaned_data.var().mean(),
    "Standard Deviation": cleaned_data.std().mean(),
}

# Step 5: Print the results
print("Statistics for Global Competitiveness Index:")
for stat, value in statistics.items():
    print(f"{stat}: {value}")
```

Figure 8: Script to sort the Global Competitive Index CSV file.

```python
import pandas as pd

# Load the dataset
file_path = '/Users/mdanylchuk/Downloads/GlobalInnovationIndex.csv'  # Replace with your file path
data = pd.read_csv(file_path)

# Step 1: Extract relevant columns (Country: 'Economy Name', Rank: 2016)
data = data[['Economy Name', '2016']]
data.columns = ['Country', 'Rank']

# Step 2: Clean the data - Drop rows with missing, non-numeric, or zero ranks
data['Rank'] = pd.to_numeric(data['Rank'], errors='coerce')
cleaned_data = data.dropna()
cleaned_data = cleaned_data[cleaned_data['Rank'] != 0]  # Ignore rows with Rank == 0

# Step 3: Sort data by Rank
sorted_data = cleaned_data.sort_values(by='Rank')

# Step 4: Calculate statistics
statistics = {
    "Mean": sorted_data['Rank'].mean(),
    "Median": sorted_data['Rank'].median(),
    "Mode": sorted_data['Rank'].mode().iloc[0] if not sorted_data['Rank'].mode().empty else None,
    "Minimum": sorted_data['Rank'].min(),
    "Maximum": sorted_data['Rank'].max(),
    "Variance": sorted_data['Rank'].var(),
    "Standard Deviation": sorted_data['Rank'].std(),
}

# Step 5: Print the sorted data and statistics
print("Sorted Data:")
print(sorted_data)
print("\nStatistics for Ranks:")
for stat, value in statistics.items():
    print(f"{stat}: {value}")
```

Figure 9: Script to sort the Global Innovation Index CSV file.

```
!pip install pandas

import pandas as pd

# URL of the shared Google Sheet with ?output=csv
google_sheet_url = "https://docs.google.com/spreadsheets/d/e/2PACX-1vR2bPb-ejxEJKleuEiOOpmwU3l3EJdKoIMihFWYhT8sVLboSLudRs4qD4UlQP_Olg/pub?output=csv"

try:
    # Read the CSV file, skipping the first 4 rows, and treating the 5th row as the header
    df = pd.read_csv(google_sheet_url, header=4)

    # Extract the 'Country Name' and numerical data columns
    df_clean = df.copy()
    df_clean['Country Name'] = df.iloc[:, 0]  # Extract the first column as 'Country Name'
    df_numeric = df_clean.iloc[:, 1:]  # Exclude the 'Country Name' column for numeric analysis

    # Convert to numeric values, replacing invalid entries with NaN (coercing errors)
    df_numeric = df_numeric.apply(pd.to_numeric, errors='coerce')

    # Calculate the mean, median, mode, max, min, variance, and std deviation for each country
    df_clean['Mean'] = df_numeric.mean(axis=1)
    df_clean['Median'] = df_numeric.median(axis=1)
    df_clean['Mode'] = df_numeric.mode(axis=1).iloc[:, 0] if not df_numeric.mode(axis=1).empty else None
    df_clean['Max'] = df_numeric.max(axis=1)
    df_clean['Min'] = df_numeric.min(axis=1)
    df_clean['Variance'] = df_numeric.var(axis=1)
    df_clean['Standard Deviation'] = df_numeric.std(axis=1)

    # Display the results for each country
    print("Country Statistics:\n", df_clean[['Country Name', 'Mean', 'Median', 'Mode', 'Max', 'Min', 'Variance', 'Standard Deviation']])

    # Calculate overall statistics (ignoring NaNs)
    combined_numeric = df_numeric.values.flatten()  # Flatten to a single array
    combined_numeric = combined_numeric[~pd.isna(combined_numeric)]  # Remove NaN values

    overall_mean = combined_numeric.mean()
    overall_median = pd.Series(combined_numeric).median()
    overall_mode = pd.Series(combined_numeric).mode().iloc[0] if not pd.Series(combined_numeric).mode().empty else None
    overall_min = combined_numeric.min()
    overall_max = combined_numeric.max()
    overall_variance = combined_numeric.var()
    overall_std = combined_numeric.std()

    # Display overall statistics
    print("\nOverall Statistics Across All Countries:")
    print(f"Mean: {overall_mean}")
    print(f"Median: {overall_median}")
    print(f"Mode: {overall_mode}")
    print(f"Minimum: {overall_min}")
    print(f"Maximum: {overall_max}")
    print(f"Variance: {overall_variance}")
    print(f"Standard Deviation: {overall_std}")

except Exception as e:
    print("Error loading or processing the data:", e)
```

Figure 10: Script to sort the Governance Indicators CSV file.